PATENT

COPY

Attorney Docket No. UCAL-250/01US
UC CASE NO. 96-152-2

METHODS FOR TREATING BIPOLAR MOOD DISORDER ASSOCIATED WITH MARKERS ON CHROMOSOME 18p

7 insert A

ACKNOWLEDGEMENTS

This invention was made with Government support under Grant Nos. RO1-MH49499, K21MH00916, awarded by the NIH. The U.S. Government has certain rights in this invention.

INTRODUCTION

Background

BIPOLAR MOOD DISORDER (BP)

Manic-depressive illness, or bipolar mood disorder (BP), is characterized by episodes of elevated mood (mania) and depression and is among the most prevalent and potentially devastating of psychiatric syndromes. The most severe and clinically distinctive forms of BP are BP-I (severe bipolar mood disorder) and SAD-M (schizoaffective disorder manic type), and are characterized by at least one full episode of mania, with or without episodes of major depression (defined by lowered mood, or depression, with associated disturbances in rhythmic behaviors such as sleeping, eating, and sexual activity). A milder form of BP is BP-II, bipolar mood disorder with hypomania and major depression. BP-I often co-segregates in families with more etiologically heterogeneous syndromes, such as unipolar major depressive disorder (MDD), which is a more broadly defined phenotype. See McInnes, L.A. and Freimer, N.B., Mapping genes for psychiatric disorders and behavioral traits, Curr. Opin. in Genet. and Develop., 5:376-381 (1995).

TREATMENT OF INDIVIDUALS WITH BIPOLAR MOOD DISORDER

An estimated 2-3 million people in the United States are affected by BP-I. Currently, individuals are typically evaluated for bipolar mood disorder using the **clinical** criteria set forth in the most current version of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM). Many drugs have been used to treat individuals diagnosed with bipolar mood disorder, including lithium salts, carbamazepine and valproic acid. However, none of the currently available drugs is able to treat every individual diagnosed with severe BP-I (termed BP-I) and drug treatments are effective in only approximately 60-70% of individuals diagnosed with BP-I. Moreover, it is currently impossible to predict which drug treatments will be effective in particular BP-I affected individuals. Commonly, upon diagnosis affected individuals are prescribed one drug after another until one is found to be effective. Early prescription of an effective drug treatment is critical for several reasons, including the avoidance of extremely dangerous manic episodes and the risk of progressive deterioration if effective treatments are not found. Also, appropriate treatment may prevent depressive episodes in BP-I individuals; these episodes are also dangerous and are characterized by a high suicide rate. The high prevalence of the disorder, together with frequent occurrence of hospitalizations, psychosocial impairment, suicide and substance abuse, has made BP-I a major public health concern.

Genetic Basis for Bipolar Mood Disorder

Mapping genes for common diseases believed to be caused by multiple genes, such as BP-I, may be complicated by the typically imprecise definition of phenotypes, by etiologic heterogeneity and by uncertainty about the mode of genetic transmission of the disease trait. With psychiatric disorders there is even greater ambiguity in distinguishing between individuals who likely carry an affected genotype from those who are genetically unaffected. For example, one can define an affected phenotype for BP by including one or more of the broad grouping of diagnostic classifications that constitute the mood disorders: BP-I, SAD-M, MDD, and BP-II.

Thus, one of the greatest difficulties facing psychiatric geneticists is uncertainty regarding the validity of phenotype designations, since clinical diagnoses are based solely on

clinical observation and subjective reports. Also, with complex traits such as psychiatric disorders, it is difficult to map the trait-causing genes genetically because: (1) the BP-I phenotype doesn't exhibit classic Mendelian recessive or dominant inheritance patterns attributable to a single genetic locus, (2) there may be incomplete penetrance i.e., individuals who inherit a predisposing allele may not manifest the disease; (3) the phenocopy phenomenon may occur, i.e., individuals who do not inherit a predisposing allele may nevertheless develop the disease due to environmental or random causes; (4) genetic heterogeneity may exist, in which case mutations in any one of several genes may result in identical phenotypes.

The existence of one or more major genes associated with BP-I and with a clinically similar diagnostic category, SAD-M (schizoaffective disorder manic subtype), is supported by segregation analyses and twin studies (Bertelson et al., 1977; Freimer and Reus, 1992; Pauls et al., 1992). However, efforts to identify the chromosomal location of BP-I genes have yielded disappointing results in that reports of linkage between BP-I and markers on chromosomes X and 11 could not be independently replicated nor confirmed in the re-analyses of the original pedigrees (Baron et al., 1987; Egeland et al., 1987; Kelsoe et al., 1989; Baron et al., 1993). The possible localization of BP genes on chromosomes 18 (pericentromeric region) and 21q has been suggested, but in both cases the proposed candidate region is not well defined and there is equivocal support for either location (Berrettini et al. (1994) Proc. Natl. Acad. Sci. USA, 91, 5918-5921, Murray, J.C., et al. (1994) Science 265, 2049-2054; Pauls et al., Am. J. Hum. Genet. 57:636-643 (1995); Maier et al., Psych. Res. 59:7-15 (1995); Straub et al., Nature Genet., 8:291-296 (1994)). Recent investigations have led to the isolation of chromosome 18-specific brain transcripts which have been suggested to be positional candidates for bipolar disorder (Yoshikawa et al., Am. J. Med. Gen. 74, 140-149 (1997)).

Despite abundant evidence that BP has a major genetic component, linkage studies have not yet succeeded in definitively localizing a BP gene. This is mainly because mapping studies of psychiatric disorders have generally been conducted under a paradigm appropriate for mapping genes for simple Mendelian disorders, namely, using linkage analysis in the expectation of finding high lod scores that definitively signpost the location of disease genes.

The follow up to early BP linkage studies, however, showed that even extremely high lod scores at a single location can be false positives. See Egeland, et al., Nature 325:783-787 (1987); Baron et al., Nature 326:289-292 (1987); Kelsoe et al., Nature, 342:238-243 (1989); and Baron et al., Nature Genet. 3:49-55 (1993). These earlier studies used largely uninformative markers and did not use stringent criteria for identifying affected individuals.

LINKAGE DISEQUILIBRIUM ANALYSIS

Linkage disequilibrium (LD) analysis is a powerful tool for mapping disease genes and may be particularly useful for investigating complex traits. LD mapping is based on the following expectations: for any two members of a population, it is expected that recombination events occurring over several generations will have shuffled their genomes, so that they share little in common with their ancestors. However, if these individuals are affected with a disease inherited from a common ancestor, the gene responsible for the disease and the markers that immediately surround it will likely be inherited without change, or IBD ("identical by descent"), from that ancestor. The size of the regions that remain shared (i.e. IBD) are inversely proportional to the number of generations separating the affected individuals and their common ancestor. Thus, "old" populations are suitable for fine scale mapping and recently founded ones are appropriate for using LD to roughly localize disease genes more approximately (Houwen et al., 1994, in particular Fig. 3 and accompanying text). Because isolated populations typically have had a small number of founders, they are particularly suitable for LD approaches, as indicated by several successful LD studies conducted in Finland (de la Chapelle, 1993).

LD analysis has been used in several positional cloning efforts (Kerem et al., 1989; MacDonald et al., 1992; Petrukhin et al., 1993; Hastbacka et al., 1992 and 1994), but in each case the initial localization had been achieved using conventional linkage methods. Positional cloning is the isolation of a gene solely on the basis of its chromosomal location, without regard to its biochemical function. Lander and Botstein (1986) proposed that LD mapping could be used to screen the human genome for disease loci, without conventional linkage analyses. This approach was not practical until a set of mapped markers covering

the genome became available (Weissenbach et al., 1992). The feasibility of genome screening using LD mapping is now demonstrated by the applicants.

Identification of the chromosomal location of a gene responsible for causing severe bipolar mood disorder can facilitate diagnosis, treatment and genetic counseling of individuals in affected families.

Due to the severity of the disorder and the limitations of a purely phenotypic diagnosis of BP-I, there is a tremendous need to subtype individuals with BP-I genetically to confirm clinical diagnoses and to determine appropriate therapies based on their genotypic subtype.

SUMMARY OF THE INVENTION

The present invention comprises using genetic linkage and haplotype analysis to identify an individual having a bipolar mood disorder gene on the short arm of chromosome 18. In addition, the present invention provides markers linked to a gene responsible for susceptibility to bipolar mood disorder that will enable researchers to focus future analysis on that small chromosomal region and will accelerate the sequencing of a bipolar mood disorder gene located at 18p.

The present invention provides, for the first time, a localization of a BP-I susceptibility locus to a 300 to 500 kb region of the short arm of chromosome 18.

The present invention is directed to methods of detecting the presence of a bipolar mood disorder susceptibility locus in an individual, comprising analyzing a sample of DNA for the presence of a DNA polymorphism on the short arm of chromosome 18 between SAVA5 and ga203, wherein the DNA polymorphism is associated with a form of bipolar mood disorder. The invention includes the use of genetic markers in the roughly 500 kb region between the SAVA5 locus and the ga203 locus, inclusive, to diagnose bipolar mood disorder genetically in individuals and to confirm phenotypic diagnoses of bipolar mood disorder. Preferably, the sample of DNA is analyzed for the presence of a DNA polymorphism on the short arm of chromosome 18 in the roughly 300 kb region between D18S1140 and W3422.

In a further embodiment, the invention provides methods of classifying subtypes of bipolar mood disorder by identifying one of more DNA polymorphisms located within the 500 kb region between SAVA5 and ga203 loci, inclusive, on the short arm of chromosome 18 and analyzing DNA samples from individuals phenotypically diagnosed with bipolar mood disorder for the presence or absence of one or more of said DNA polymorphisms. Preferably, the sample of DNA is analyzed for the presence or absence of one or more of said DNA polymorphisms in the roughly 300 kb region between D18S1140 and W3422 on the short arm of chromosome 18.

In yet a further embodiment, the methods of the invention include a method of treating an individual diagnosed with bipolar mood disorder comprising identifying one or more DNA polymorphisms located within the 500 kb region of chromosome 18 between SAVA5 and ga203, analyzing DNA samples from individuals phenotypically diagnosed with bipolar mood disorder for the presence or absence of one or more of the DNA polymorphisms, and selecting a treatment plan that is most effective for individuals having a particular genotype within the 500 kb region of chromosome 18 between SAVA5 and ga203. Preferably, the sample of DNA is analyzed for the presence or absence of one or more DNA polymorphisms in the roughly 300 kb region between D18S1140 and W3422 on the short arm of chromosome 18.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a pedigree chart showing two families, CR001 and CR004. Affected individuals are denoted by black symbols, deceased individuals by a diagonal slash. A schematic of each individual's haplotype (where available) is shown below the ID number. Recombinations are denoted by "-x"; consanguineous marriages by a double bar, and the conserved haplotype as black shading within the haplotype bars. The larger conserved region for CR004 is stippled, the larger conserved region for CR001 is indicated by a dashed outline. An "I" underneath the haplotype bars indicates inferred haplotype. A "?" indicates phase is uncertain. The connection between CR001 and CR004, dating to an 18th Century founding couple, is indicated by the dashed lines joining individuals III-6 and I-4.

FIG. 2 is a table of lod scores for markers covering the entire human genome that exceeded the arbitrary coverage thresholds. Lod scores are shown for two markers on chromosome 18: D18S59 and D18S1105.

FIG. 3 depicts the extent of marker coverage used in the pedigree genome screening study for each chromosome. Coverage is defined as regions for which a lod score of at least 1.6 would have been detected (in the combined data set) for markers truly linked to BP-I under the model employed. Areas that remain uncovered (at this threshold) are unshaded. Markers for which lod scores were obtained that exceeded the empirically determined coverage thresholds in CR001, CR004, or the combined data set, are shown at their approximate chromosomal location. The symbols to the right of the chromosome indicate the thresholds exceeded at that marker: a circle signifies that the lod score at a marker exceeded the threshold of 0.8 in CR001, a diamond signifies that the lod score exceeded the threshold of 1.2 in CR004, and a star signifies that the lod score exceeded the threshold of 1.6 in the combined data set.

FIGS. 4A and 4B depicts the Lod score for the maximum likelihood estimate of theta in the combined sample for the 473 microsatellite markers typed in the pedigree genome screen. The MLEs of theta were appointed to the following categories: theta < 0.10; $0.10 \leq$ theta ≤ 0.40; theta ≥ 0.40. Note that the scale for the x-axis (distance from pter) changes with chromosomes.

FIG. 5 is a portion of an integrated map of the 5 cM 18pter region of chromosome 18.

FIGS. 6A, 6B and 6C are a list of markers on chromosome 18, with map positions noted.

FIG. 7 describes 18p allele frequencies for disease chromosomes (aff 105) versus nontransmitted chromosomes (ntrans) and samples from a control population of Costa Rican

students and their parents (control). The name of each marker used in this study is indicated on the left. The second column of numbers refers to allele length in base pairs.

FIG. 8 depicts haplotype analysis of individuals affected with BP-I. The column labelled 18p refers to the patient identifier, and each patient identifier is repeated with 2 rows to indicate allele results with each of the patient's two copies of chromosome 18. The columns labelled "PANR" and "MANR" refer to the paternal and maternal identifiers, respectively, associated with the particular patient, other than 0, 1 and 2, which indicate that parental samples were not available. The column headings to the right of "PANR" and "MANR" columns represent names of specific markers in the 18p region that were used in the haplotype analysis. The markers are listed in the order they appear on chromosome 18. The allele length (in base pairs) is indicated under the column heading each marker for a particular patient. In the column to the immediate right of each marker column, a "1" indicates that the phase is known, i.e., that it is known whether a particular allele is inherited from the paternal or maternal chromosome, and a "0" indicates that the phase is not definitely known. The shaded horizontal bars depict haplotypes of at least three markers which include a 154 allele length at D18S59, other than patients 218, 225, 232, 234, 311, 314 and 458, where the stippled region depicts small sections that do not have the 154 allele at D18S59. The hatched regions depict uncertainty as to whether the individual has the affected haplotype, as the phase is not known with certainty. In addition, the presence of an allele length of 232 (or 234) with marker ta201 is thought to result from a highly mutable allele and may not be distinct from the 230 allele. Similarly, the 202 allele at ca212 may not be distinct from the 200 allele at ca212. Patients 246, 247, 248, 311, 316, 367, 384, 501, 531, 587, 536, 684, 667 and 669 exhibit a 242, 244, 250, 252 or 214 allele at marker ta201 which indicates a potential marker location. Patients 488, 435 and 236 exhibit haplotypes that are distinct from the pedigrees that were analyzed.

FIG. 9 depicts haplotype analysis of nontransmitted chromosomes from parents of individuals affected with BP-I. The labels "ERSN" and "KID" refer to the parental and patient identifiers, respectively. As above, allele length is provided in base pairs below each

marker with an indication as to whether phase was known (1) or not known (0) given to the right of these values. The markers, shading and allele characteristics described for Figure 8 also apply to this figure.

FIG. 10 depicts haplotype analysis of control samples obtained from an unscreened population of students of the University of Costa Rica and their parents representing the general population. Identifiers are provided in the column headed "cont", allele length and phase determination given in the remainder of the table. The markers, shading and allele characteristics described for Figure 8 also apply to this figure. Complete data for all markers are not given as indicated by blank boxes, or the terms "miss" or "missing".

FIG. 11 depicts Ancestral Haplotype Reconstruction results in disease chromosomes.

DESCRIPTION OF SPECIFIC EMBODIMENTS

The recent availability of highly polymorphic, genetically mapped markers covering the human genome (Weissenbach, J., et al. (1996) Nature 359, 794-801, Murray, J.C., et al. (1994) Science 265, 2049-2054, Gyapay, G., et al. (1994) Nature Genet 7,246-339) has allowed the development of a multi-stage paradigm for mapping genes for complex traits. In the first stages, complete genome screening (e.g. through lod score analysis) is used to identify possible localizations for disease genes. Subsequently, the regions highlighted by the screening study are more intensively investigated to confirm the initial localizations and delineate clear candidate regions. Finally, fine mapping methods (such as haplotype or linkage disequilibrium (LD) analysis) or candidate gene approaches are used for positional cloning of disease genes.

Our genome screening study for BP employed the following strategies. Unlike previous genetic studies of BP, only those individuals with the most severe and clinically distinctive forms of BP (BP-I and schizoaffective disorder manic type, SAD-M) were considered as affected, rather than including those diagnosed with a milder form of BP (BP-II) or with unipolar major depressive disorder (MDD). Two large pedigrees (CR001 and

CR004) were selected from a genetically homogeneous population, that of the Central Valley of Costa Rica (as described in Escamilla, M.A., et al., (1996) Neuropsychiat. Genet. 67, 244-253, and in Freimer, N.B., et al. (1996) Neuropsychiat. Genet. 67, 254-263, both incorporated by reference herein). The entire human genome was screened for linkage using mapped microsatellite markers and a model for genetic analysis in which most of the linkage information was derived from affected individuals. The goal of this stringent linkage analysis was to identify all regions potentially harboring major genes for BP-I in the study population. Empirically determined lod score thresholds (using linkage simulation analyses) were derived, to suggest regions worthy of further investigation.

Identification of all suggestive regions and weighing the relative importance of findings required complete screening of the genome. The coverage approach was developed to gauge the progress of this effort. Conventionally, the thoroughness of genome screening is evaluated by excluding genome regions from linkage under given genetic models. This approach, which is highly sensitive to misspecification of genetic models, may be poorly suited for genome screening studies of complex traits; it is tied to the expectation of finding linkage at a single locus and demonstrating absence of linkage at all other locations in the genome. Additionally, exclusion analyses do not differentiate between genome regions where linkage is not excluded because markers are uninformative in the study population from those in which the genotype data are simply ambiguous. In contrast, the coverage approach is designed for studies aimed at genome screening rather than for studies where the goal is to demonstrate a single unequivocal linkage finding, and it provides explicit data regarding the informativeness of markers in the study pedigrees. Its use lessens the possibility that one would prematurely dismiss a given genome region as being unpromising for further study.

Because the exact genetic length of chromosomes is not clearly established, it is impossible to be certain that one has screened the entire genome. Although we report coverage of about 94% of the genome (under the 90%) dominant model) at the thresholds described above, this probably represents an underestimate. The remaining coverage gaps in our study occur predominantly at or near telomeres; as the upper bound estimates for the

length of each chromosome were used, it is likely that the actual coverage gaps in these regions are smaller than our conservative assessment.

The presence of consistently positive lod scores over a given region was considered to be of greater significance than isolated peak lod scores. Such clustering suggests true co-segregation of markers and phenotypes (i.e. alleles are shared identically by descent rather than identically by state) and is more readily observed in analyses of a few large pedigrees (as in our study) than in examination of several smaller families. The data presented herein indicates clustering of positive lod scores in the region of the telomere of 18p.

The genome screen was conducted in two stages. The Stage I screen identified areas suggestive of linkage, so that those areas could be saturated with available markers, and so that regions, referred to as 'coverage gaps', could be pinpointed where markers were insufficiently informative in our sample to detect evidence of linkage. The Stage II screen followed up on regions flanking each marker that yielded peak lod scores approximately equal to or greater than the thresholds used for the coverage calculations, which were deemed regions of interest, and filled in coverage gaps. The results of the complete genome screen (Stages I and II) using 473 markers is described below.

In addition, linkage disequilibrium analysis of an independently collected sample of 48 unrelated BP-I patients was initially conducted. These patients were from the same ancestral population as the patients in the CR001 and CR004 pedigrees. The LD analysis was conducted with markers on the short arm of chromosome 18 (18p), in a 5 centimorgan (cM) region ("5 cM 18pter region") extending from the end of the 18p telomere to a distance of 5 cM along the short arm of chromosome 18 (18p). The LD analysis gave evidence of LD in this region, particularly at marker D18S59 and also at D18S476. LD analysis of further BP-I patients from the CRCV with markers in this 5 cM 18pter region was conducted to confirm and fine map a BP-I gene in this region. This approach, using additional BP-I patients from this CRCV population and additional markers identifies the region of maximum LD and can precisely localize a BP-I susceptibility gene.

Fine mapping of 5 cM 18pter region resulted in the identification of two DNA markers (D18S1140 and W3422) defining the boundaries of BP-I as approximately 300 kb, thus allowing a systematic search for the BP-I gene(s).

A conservative approach to linkage analysis was used in that almost all of the information for linkage is derived from individuals with a severe, narrowly defined phenotype. While this approach made it very unlikely that lod scores greater than conventional thresholds of statistical significance (e.g. ≥ 3) would be obtained, it provided confidence in the robustness of the most suggestive findings.

Direct cDNA selection can be used to isolate segments of expressed DNA from the 300 kb region between D18S1140 and W3422 (M. Lovett, J. Kere, L.M. Hinton, *Proc. Natl. Acad. Sci. USA* 88 9628-9632 (1991); Y.-S. Jou *et al., Genomics* 24 410-413 (1994)). By using bacterial artificial chromosomes (BAC) (e.g., commercially available from Research Genetics Inc. Huntsville, Alabama), a group of cDNAs can be identified, and hybridization and PCR-amplification experiments can be used to determine if these cDNA segments are derived from the 300 kb region.

The cDNAs can then be used to determine whether specific sequences are expressed at lower levels (or not at all) in affected individuals compared to non-carrier individuals. Measurement of mRNA levels in lymphoblastoid cell lines can be used as an initial screen. The cell lines are prepared by drawing blood from individuals, transforming the lymphoblasts with EBV and growing the immortalized cells in culture. Total RNA and DNA are extracted from the cultured human lymphoblastoid cell lines. Northern blot hybridization is used to determine reduced levels of a specific sequence compared to levels from an unaffected, non-carrier individual as a result of mutations in the BP-I gene on the chromosomes from these affected individuals which results in decreased levels of mature mRNA and play a primary role in BP-I. Thus, alterations in gene sequences in affected individuals can be determined.

The polymerase chain reaction (PCR) is used to amplify the gene and to determine its sequence from affected individuals. Sequence comparison with unaffected, non-carrier individuals is carried out to identify polymorphisms in the gene sequence that are responsible for BP-I.

The identification of the biochemical defect that causes BP-I provides a basis for treatments for this disease. In addition, knowledge that certain mutations in the gene are responsible for the disease allows mutation detection tests to be used as a definitive diagnosis for BP-I.

Thus, the present invention allows the isolation of a nucleic acid molecule that can be used in the identification of the presence (or absence) of a mutation in the BP-I gene a human and thus can be used in the diagnosis of BP-I or in the genetic counseling of individuals, for example those with a family history of BP-I (although the general population can be screened as well). In particular, it should be noted that any mutation in the BP-I gene away from the normal gene sequence is an indication of a potential genetic flaw; even so-called "silent" mutations that do not encode a different amino acid at the location of the mutation are potential disease mutations, since such mutations can introduce into (or remove from) the gene an untranslated genetic signal that interferes with the transcription or translation of the gene. Thus, advice can be given to a patient concerning the potential for transmission of BP-I if any mutation is present. While an offspring with the mutation in question may or may not have symptoms of BP-I, patient care and monitoring can be selected that will be appropriate for the potential presence of the disease; such additional care and/or monitoring can be eliminated (along with the concurrent costs) if there are no differences from the normal gene sequence. As additional information (if any) becomes available (e.g., that a given silent mutation or conservative replacement mutation does or does not result in BP-I), the advice given for a particular mutation may change. However, the change in advice given does not alter the initial determination of the presence or absence of mutations in the gene causing BP-I.

Generally, mutations are identified in the human gene for use in a method of detecting the presence of a genetic defect that causes or may cause BP-I, or that can or may transmit BP-I to an offspring of the human. Initially, the practitioner will be looking simply for differences from the sequence identified as being normal and not associated with disease, since any deviation from this sequence has the potential of causing disease, which is a sufficient basis for initial diagnosis, particularly if the different (but still unconfirmed) gene is found in a person with a family history of BP-I. As specific mutations are identified as being positively correlated with BP-I (or its absence), practitioners will in some cases focus on identifying one or more specific mutations of the gene that changes the sequence of a protein product of the gene or that results in the gene not being transcribed or translated. However, simple identification of the presence or absence of any mutation in the gene of a

patient will continue to be a viable part of genetic analysis for diagnosis, therapy and counseling.

The actual technique used to identify the gene or gene mutant is not itself part of the practice of the invention. Any of the many techniques to identify gene mutations, whether now known or later developed, can be used, such as direct sequencing of the gene from affected individuals, hybridization with specific probes, which includes the technique known as allele-specific oligonucleotide hybridization, either without amplification or after amplification of the region being detected, such as by PCR. Other analysis techniques include single-strand conformation polymorphism (SSCP), restriction fragment length polymorphism (RFLP), enzymatic mismatch cleavage techniques and transcription/translation analysis. All of these techniques are described in a number of patents and other publications; see, for example, "Laboratory Protocols for Mutation Detection" (1996) Oxford University Press, Editor: Ulf Landegrun.

Depending on the patient being tested, different identification techniques can be selected to achieve particularly advantageous results. For example, for a group of patients known to be associated with particular mutations of the gene, oligonucleotide ligation assays, "mini-sequencing" or allele-specific oligonucleotide (ASO) hybridization can be used. For screening of individuals who are not known to be associated with a particular mutation, single-strand conformation polymorphism, total sequencing of genetic and/or cDNA and comparison with standard sequences are preferred.

In many identification techniques, some amplification of the host genomic DNA (or of messenger RNA) will take place to provide for greater sensitivity of analysis. In such cases it is not necessary to amplify the entire gene, merely the part of the gene or the specific location within the gene that is being detected. Thus, the method of the invention generally comprises amplification (such as via PCR) of at least a segment of the gene, with the segment being selected for the particular analysis being conducted by the diagnostician.

The patient on whom diagnosis is being carried out can be an adult, as is usually the case for genetic counseling, or a newborn, or prenatal diagnosis can be carried out on a fetus. Blood samples are usually used for genetic analysis of adults or newborns (e.g.,

screening of dried blood on filter paper), while samples for prenatal diagnosis are usually obtained by amniocentesis or chorionic villus biopsy.

Prior to the present invention, affected individuals were prescribed one drug after another until one was found to be effective. As BP-I was diagnosed using clinical criteria, no correlation between using a particular drug and its efficacy in a given case was observed. As a result of the present invention, BP-I subtypes can be diagnosed at the molecular level and effective treatment predicted.

For example, lithium salts, carbamazepine and valproic acid have all been prescribed for BP-I affected individuals with serendipitous results. An individual can now be diagnosed with bipolar mood disorder by analyzing genetic material from that individual for the presence or absence of one or more nucleic acid mutations as described above. As a result of this diagnosis at the molecular level, an effective treatment can be determined by collecting data to obtain a statistically significant correlation of a particular treatment with the different subtypes of BP-I. Thus, the practitioner is able to select a specific drug for the treatment of a particular sub-type of BP-I and does not merely rely on trial and error.

Alternatively, the full-length normal genes for BP-I from humans, as well as shorter genes that produce functional proteins, can be used to correct BP-I in a human patient by supplying to the human an effective amount of a gene product of the human gene, either by gene therapy or by *in vitro* production of the protein followed by administration of the protein. It should be recognized that the various techniques for administering genetic materials or gene products are well known and are not themselves part of the invention. The invention merely involves supplying the genetic materials or proteins identified as a result of the present invention in place of the genetic materials or proteins previously administered. For example, techniques for transforming cells to produce gene products are described in U.S. Patent No. 5,283,185 entitled "Method for Delivering Nucleic Acid into Cells," as well as in numerous scientific articles, such as Felgner et al., "Lipofection: A Highly Efficient, Lipid-Mediated DNA-Transfection Procedure," *Proc. Natl. Acad. Sci. U.S.A.*, 84, 7413-7417 (1987); techniques for in vivo protein production are described in, for example, Mueller et al., "Laboratory Methods - Efficient Transfection and Expression of Heterologous Genes in PC12 Cells," *DNA and Cell Biol.*, 9(3), 221-229 (1990).

Administration of proteins and other molecules to overcome a deficiency disease is well known (e.g., administration of insulin to correct for high blood sugar in diabetes) that further discussion of this technique is not necessary. Some modification of existing techniques may be required for particular applications, but those modifications are within the skill level of the ordinary practitioner using existing knowledge and the guidance provided in this specification.

The invention now being generally described, the following examples are provided for purposes of illustration only and are not to be considered to limit the invention.

EXAMPLES

PEDIGREES

Two independently ascertained Costa Rican pedigrees (CR001 and CR004) were chosen because they contained a high density of individuals with BP-I and because their ancestry could be traced to the founding population of the Central Valley of Costa Rica. The current population of the Central Valley (consisting of about two million people) is predominantly descended from a small number of Spanish and Amerindian founders in the 16th and 17th centuries (Escamilla, M.A., et al., (1996) Neuropsychiat. Genet. 67, 244-253). Studies of several inherited diseases have confirmed the genetic isolation of this population (Leon, P., et al. (1992) Proc. Natl. Acad. Sci. USA. 89, 5181-5184; Uhrhammer, N., et al. (1992) Am. J. Hum. Genet. 57, 103-111). An extensive description of pedigrees CR001 and CR004 has ben published (Freimer, N.B., et al. (1996) Neuropsychiat. Genet. 67, 254-263). In the course of the study, two links between these pedigrees were discovered. However, the families were analyzed separately because these links were discovered after the simulation analyses were completed and after the genome screening study had been initiated.

All available adult members of these families were interviewed in Spanish using the Schedule for Affective Disorders and Schizophrenia Lifetime version (SADS-L) (Endicott, J. et al, (1978) Arch. Gen. Psych. 35, 837-844). Individuals who received a psychiatric diagnosis were interviewed again in Spanish by a research psychiatrist using the Diagnostic

Interview for Genetic Studies (DIGS) (Nurnberger, J.L. et al. (1994) Arch. Gen. Psychiat. 51, 849-859). This recently developed instrument is similar to, but more detailed than SADS-L. The interviews and medical records were then reviewed by two blinded best estimators who reached a consensus diagnosis. The diagnostic procedures are described in detail in Freimer, N.B., et al. (1996) Neuropsychiat. Genet. 67, 254-263 (incorporated by reference herein).

UNRELATED CRCV BP-I PATIENT STUDY

BP localizations obtained through the CRCV pedigree studies were confirmed by genotyping an independently collected sample of 48 unrelated BP-I patients from the CRCV. In this fine mapping LD analysis, 48 unrelated BP-I patients from the CRCV were identified and genotyped using microsatellite markers spaced at narrow intervals across chromosome 18. As these patients are descended from the same ancestral population as the patients in the pedigrees previously studied (CR001 and CR004), many of them should share disease susceptibility alleles inherited identically by descent (IBD) from one or a few common ancestors, and linkage disequilibrium (LD) should be present at marker loci surrounding the disease genes.

The sample of 48 BP-I patients included 25 women and 23 men who were recruited from psychiatric hospitals and clinics in the CRCV. These patients were ascertained only on the basis of diagnosis and CV ancestry, and were not selected on the basis of history of BP illness in family members. A structured interview of each patient was conducted by a psychiatrist, and medical and hospital records were collected. Ascertainment and diagnostic procedures were as described above. However, in order to lessen further the probability of phenocopies among this unrelated sample, for which we lacked pedigree information, the affected phenotype was defined even more narrowly than in the pedigree study. Individuals considered affected in this study had to have suffered at least two disabling episodes of mania (requiring hospitalization) and a first onset of the illness before age 45.

Genealogical research on each of the 48 BP-I patients confirmed that on average, 70% of their great-grandparents were born in the CRCV. Individuals whose great-grandparents were born in the CRCV were considered likely to be descended from the original Spanish

and Amerindian founders of the CRCV. Genealogical research showed that 2 patients are first cousins and the remaining 46 have no relationship within the past 4 generations.

GENOTYPING PEDIGREE STUDIES

Linkage simulations were used to select the most informative individuals from pedigrees CR001 and CR004 for genotyping studies (Freimer, N.B., et al. (1996) Neuropsychiat. Genet. 67, 254-263). Under a 90% dominant model, simulation analyses with these individuals suggested that evidence of linkage would likely be detected (e.g. a probability of 92% of obtaining lod > 1.0 in the combined data set) using markers with an average heterozygosity of 0.75 spaced at 10 cM intervals (as discussed in Freimer, N.B., et al. (1996) Neuropsychiat. Genet. 67, 254-263). For the Stage I screen, the most polymorphic markers (307 in total) were chosen, placed at approximately 10 cM intervals on the 1992 Genethon map (Houwen, R., et al. (1992) Nature 359, 794-801). These markers were then supplemented by a small number of markers from the Cooperative Human Linkage Center (CHLC) public database. For the Stage II screen, 166 markers were added from newer Genethon and CHLC maps as they became available (Murray, J.C. et al. (1994) Science 265, 2049-2054, Gyapay, G., et al. (1994) Nature Genet. 7,246-339) and from the public database of the Utah Center for Genome Research, all of which are publicly available. DNA samples (from individuals in the CEPH families) that were used for size standards for Genethon and CHLC markers were included in the experiments to permit comparison of allele sizes between members of the CRCV population and individuals in the CEPH database. Genotyping procedures were as described previously (DiRienzo, A. et al. (1994) Proc. Natl. Acad. Sci. USA 91, 3166-3170 (incorporated by reference herein)). Briefly, one of the two PCR primers was labeled radioactively using a polynucleotide kinase and PCR products were run on polyacrylamide gels. Autoradiographs were scored independently by two raters. Data for each marker were entered into the computer database twice and the resultant files were compared for discrepancies.

GENOTYPING OF UNRELATED BP-I CRCV PATIENTS

Twenty-seven markers were used to genotype all 48 individuals (as well as 53 individuals used to establish genetic phase) at approximately 5 cM intervals along the entire chromosome 18. It was hypothesized that such a screen would permit the evaluation of evidence in the 18pter region and also to investigate other regions on chromosome 18 in which linkage to BP has been suggested by other groups in other populations. For each individual, two-marker haplotypes in each of the 26 inter-marker intervals were investigated. For 38 of the 48 BP-I patients, genotypes of parents or children were available to assist in phase determination. Because of phase ambiguities in the remaining 10 individuals, minimal and maximal two-marker haplotype sharing was evaluated as follows: (1) Minimal: the number of individuals (and chromosomes) who definitely shared a chromosomal segment defined by a particular pair of alleles (phase known chromosomes) and (2) Maximal: the number of individuals (and chromosomes) who could possibly share a chromosomal segment defined by a particular pair of alleles (includes phase unknown chromosomes). The threshold used to identify areas of high IBD sharing of chromosomes in this initial screen was designated as maximal sharing of a two-marker haplotype by 50% or more of the 48 individuals (or 25% or more of the 96 chromosomes).

Arbitrary thresholds were designated to identify possible areas of high IBD sharing among the 48 patients. Eight of the 26 regions passed this screen. Within each of these 3 regions, one to three additional markers were typed to permit detection of LD, if present, over regions of one to two cM.

A total of 42 chromosome 18 markers were used to genotype the study sample: D18S1140, D18S59, D18S476, D18S481, D18S391, D18S452, D18S843, D18S464, D18S1153, D18S378, D18S53, D18S453, D18S40, D18S66, D18S56, D18S57, D18S467, D18S460, D18S450, D18S474, D18S69, D18S64, D18S1134, D18S1147, D18S60, D18S68, D18S55, D18S477, D18S61, D18S488, D18S485, D18S541, D18S870, D18S469, D18S874, D18S380, D18S1121, D18S1009, D18S844, D18S554, D18S461, D18S70 (from pter to qter). Of these 42 markers, four are located within the 5 cM 18pter region extending from the telomere of 18p to marker D18S481 (inclusive), which is approximately 5 cM from the

telomere of 18p. This region is referred to as the 5 cM 18pter region. The four markers tested in the 5 cM 18pter region are: D18S59, D18S1140, D18S476 and D18S481.

For each marker the likelihood that a particular allele (or alleles) is over-represented on disease chromosomes, as compared to non-disease chromosomes was evaluated. The results of this likelihood test provide a conservative but powerful measure of LD between two loci.

PEDIGREE STATISTICAL ANALYSES

Two-point linkage analyses were performed for all markers. Marker allele frequencies were estimated from the combined data set with correction for dependency due to family relationships (Boehnke, M. (1991) Am. J. Hum. Genet. 48, 22-25). The linkage analyses for Stages I and II included the 65 individuals who were genotyped as well as an additional 65 individuals who had been diagnostically evaluated but not genotyped. Only individuals with BP-I were considered affected with the exception of two persons, one in each family, who carry diagnoses of schizoaffective disorder manic type (SAD-M). The SAD-M individuals were included as affected because BP-I and SAD-M are often difficult to distinguish from each other based on their clinical presentation and course of illness (Goodwin, F.K. et al. (1990) in Manic Depressive Illness (Oxford University Press, New York), pp. 373-401; Freimer, N.B et al. (1993) in The Molecular and Genetic Basis of Neurological Disease, pp. 951-965; Freimer, N.B. et al. (1996) Neuropsychiat. Genet. 67, 254-263; and Freimer, N.B. et al (1996) Nature Genetics 12:436-441, all incorporated by reference herein). In all, 20 individuals were designated as affected within CR004 (Copeman, J.B., et al. (1995) Nature Genet. 9, 80-85 available for genotyping) and 10 individuals from CR001 (Kelsoe, J.R. et al. (1989) Nature 342, 238-243 available for genotyping). The phenotype for all other individuals was designated as unknown except for 17 individuals who were designated as unaffected because they had been thoroughly clinically evaluated, showed no evidence of any psychiatric disorder, and were well beyond the age of risk (50) for BP-I (linkage simulation studies indicated that these unaffected individuals contributed little information to the linkage analysis).

Linkage analyses were performed using a nearly dominant model (assuming penetrance of 0.81 for heterozygous individuals of 0.9 for homozygotes with the disease mutation). This model was chosen from five different single-locus models (ranging from recessive to nearly dominant) due to its consistency with the segregation patterns of BP in the two pedigrees and because it had demonstrated the greatest power to detect linkage in simulation studies (Freimer, N.B., et al. (1996) Neuropsychiat. Genet. 67, 254-263). Based on Costa Rican epidemiological surveys Escamilla, M.A., et al., (1996) Neuropsychiat. Genet. 67, 244-253, the population prevalence of BP-I was assumed to be 0.015 (and thus the frequency of the disease allele was assumed to be 0.003)(based on epidemiological surveys in Costa Rica, Adis, G. (1992) "Disordenes mentales en Costa Rica: Observaciones Epidemiologicas," (San Jose, Costa Rica: Editorial Nacional de Salud y Seguridad Social)). The frequency of BP-I in individuals without the disease allele was conservatively set at 0.01 which effectively specified a population phenocopy rate of 0.67 (i.e., an affected individual in the general population has a 2/3 probability of being a phenocopy). For multiply affected families, the probability that a gene segregates is highly increased, which implies that affected individuals in our study pedigree have a lower probability to be phenocopies than affected individuals in the general population, particularly those with several affected close relatives (the exact probabilities are dependent on the degree of relationship between patients and the number of intervening unaffected individuals). These parameters were chosen to ensure that most of the linkage information derives from affected individuals. The rationale for selecting these parameters and results of analyses that demonstrate the conservatism of this model are described by Freimer, N.B., et al. (1996) Neuropsychiat. Genet. 67, 254-263. The LINKAGE package (Lathrop et al., (1984) Proc. Natl. Acad. Sci. USA 81, 3443-3446) was used for lod score analysis and to obtain maximum likelihood estimates of the marker allele frequencies, taking into account the existing family relationships (see Boehnke, Am. J. Hum. Gent. 48, 22-25 (1991)).

UNRELATED BP-I CRCV PATIENT STATISTICAL ANALYSES

A likelihood test of disequilibrium (J. Terwilliger, Am. J. Hum. Genet. 56, 777 (1995)) was used to estimate a single parameter, lambda, that quantifies the over-

representation of marker alleles on disease chromosomes as compared to non-disease chromosomes. We chose this method of analysis over another commonly used disequilibrium analysis method, the transmission disequilibrium test (TDT, R. Spielman et al., Am. J. Hum. Genet. 52, 506 (1993)) because data from all 48 BP-I patients could be used in the likelihood approach. Effective use of the TDT requires phase-known, heterozygous parental chromosomes. We do not have parental genotypes for 20 of the 48 BP-I patients. Simulations indicated that with our data, the likelihood test of disequilibrium would be more powerful than the TDT. Lambda has been shown to be a superior measure for LD fine mapping, compared to other frequently used measures, because it is directly related to the recombination fraction between the disease and the marker loci. Non-disease chromosomes were chosen from the phase-known chromosomes of parents, spouses and children of affected individuals, if available. Designation of chromosomes of family members as non-disease in a disorder such as BP-I, which is not fully penetrant, necessitates specifying a model of disease transmission. The same model of transmission was employed in this LD likelihood test as was used in the initial genome screen of the pedigrees CR001 and CR002 described herein. One parameter was specified differently from the genome screen: the phenocopy rate was set to zero in the LD likelihood analysis. A phenocopy rate was not specified in the transmission model because the effect of phenocopies will be "absorbed" by the lambda parameter, in that presence of phenocopies in our sample will serve to erode the association between marker alleles and disease, and hence reduce the estimate of lambda.

COVERAGE

To access coverage for a marker, the number of informative meioses at the estimated recombination fraction was calculated using the estimate of the variance (the inverse of the information matrix) (Petrukhin, K.E. et al. (1993) Genomics 15, 76-85). Alternatively, when the estimated frequency of recombination was close to 0 or 1, Edwards' equation was applied to calculate the equivalent number of observations (Edwards, J.H. (1971) Ann. Hum. Genet. 34, 229-250). These meioses represent the amount of linkage information provided by the marker, given the pedigree structure and the genetic model applied. Linkage to the

marker in question was then assumed and the lod score that would be observed as a disease gene is hypothetically moved in increments away from that marker was calculated. All regions around a marker that would have generated a lod score that exceeded our thresholds for possible linkage (0.8 in CR001, 1.2 in CR004, and 1.6 in the combined data) were considered covered. These lod score thresholds were derived from simulation analyses showing the expected distribution of lod scores under linkage and non-linkage (Freimer, N.B., et al. (1996) Neuropsychiat. Genet. 67, 254-263, and approximately represent a result that is 250 times more likely to occur in linked simulations than in unlinked simulations. Coverage maps were constructed (**FIG. 1**) by superimposing the regions covered by each marker on the genetic map of each chromosome. At the end of the Stage II screen, a total of 473 microsatellite markers had been typed with genome coverage (in the combined data set) of over 94%. Possible coverage gaps are indicated by unshaded areas and are mainly concentrated near telomeres. Because the coverage calculations make use of marker informativeness within the pedigrees, the coverage approach thus permits detection of instances where markers with expected high heterozygosities are uninformative in our data set.

PEDIGREE LINKAGE ANALYSIS RESULTS

Of the 473 microsatellites analyzed with two-point linkage tests, 23 markers exceeded the empirically determined thresholds designated for the coverage calculations (in either CR001, CR004, or in the combined data set). The location of these markers, the peak lod scores obtained in each family and in the combined data set, and the maximum likelihood estimate of the recombination fraction (0) at which these lod scores were observed are indicated in **Table 1**. The approximate chromosomal locations of these markers are also depicted in **FIG. 1**. The distribution of lod scores (for the maximum likelihood estimate of 0 in the combined data set) across the genome is displayed by chromosome in **FIG. 2**.

The threshold was exceeded for pedigree CR001 in two adjacent markers near the 18p telomere (D18S59 and D18S1105), but CR004 displayed no suggestion of linkage in this region.

In the genome screen, the highest lod score observed for family CR001 alone was at D18S59 (1.32 at θ=0.0), located near pter. All affected members of CR001 shared alleles at markers in the 18pter region.

UNRELATED BP-I CRCV PATIENT STUDY RESULTS

Out of the forty-two markers tested, eight displayed evidence of over-representation of a particular allele on disease chromosomes. Eight of the 42 markers had -2*ln(likelihood ratio) statistics >1.0. Three other markers had -2*ln(likelihood ratio) statistics >0 and <0.62. The results are shown in Table I:

Table I

Marker	Allele Size	Frequency on non-disease Chromosomes	Frequency on Disease Chromosomes
D18S59	154	0.121	0.572
D18S476	271	0.470	0.771
D18S467	172	0.384	0.693
D18S61	177	0.074	0.326
D18S485	182	0.237	0.586
D18S870	179	0.405	0.657
D18S469	234	0.128	0.450
D18S1121	168	0.171	0.553

Evidence for association was found at markers located near the telomere of the short arm of chromosome 18. D18S59 displayed the strongest evidence for LD (-2*ln(likelihood ratio) of 8.3, p=0.002) of all the chromosome 18 markers tested. An adjacent marker, D18S476 (-2*ln(likelihood ratio) of 1.3), also provided evidence of LD. In our genome screening pedigree study we observed the single highest lod score for pedigree CR001 of any marker in the entire genome at D18S59. Furthermore, the alleles at D18S59 and D18S476

that are over-represented among the BP-I patients from the population sample (154 b.p. and 271 b.p. respectively) are observed in all BP-I patients from pedigree CR001.

The LD and pedigree findings in the 5 cM 18pter region denote a clearly delineated region that contains a BP-I susceptibility locus. This region is distinct from other regions on chromosome 18 that have been suggested as linked to mood disorder phenotypes (more broadly defined than BP-I). See **FIG. 6A, 6B, 6C.** In contrast to previous reports by Berrettini et al. and Stine et al., suggesting possible linkage between mood disorder and markers in the pericentromeric region of chromosome 18, our results did not show any evidence for association of BP-I with any pericentromeric markers (D18S378, D18S53, D18S453 or D18S40).

IDENTIFICATION OF NEW MARKERS FROM THE 5 CM 18PTER REGION

Cloned human genomic DNA covering the target region is assembled. Microsatellite sequences from these clones are identified. A sufficient area around the repeat to enable development of a PCR assay for genomic DNA is sequenced, and it is confirmed that the microsatellite sequence is polymorphic, as several uninformative microsatellites are expected in any set. Several methods have been routinely used to identify microsatellites from cloned DNA, and at this time no single one is clearly preferable (Weber, 1990, Hudson et al., 1992). Most of these require screening an excessive number of small insert clones or performing extensive subcloning using clones with larger inserts.

New strategies have recently been developed which permit the use of the several different microsatellites to be found within a single large insert clone without requiring extensive subcloning. A method for direct identification of microsatellites from yeast artificial chromosomes (YACs) provides several new markers from the target region. This procedure is based on a subtractive hybridization step that permits separation of the target DNA from the vector background. This step is useful because the human DNA (the YAC) constitutes only a small proportion of the total yeast genomic DNA.

YAC clones (with inserts averaging about 750 Kb of human genomic DNA) that span the 5 cM 18pter region have already been identified by the CEPH/Généthon consortium (Cohen et al., 1993) and are publicly available. The markers from YACs that have been

mapped to portions of the candidate region that are not well represented by currently available markers are first isolated. By typing these markers in the families and the "LD" sample, as described above, it is possible to narrow the candidate region, perhaps to a size of less than one to two cM, thus permitting limitation of the segment in which more extensive mapping efforts are applied.

Briefly, the microsatellite identification procedure is performed as follows: A subtractive hybridization is performed using genomic DNA from a target YAC together with an equivalent amount of a control DNA. This procedure separates the YAC DNA from that of the yeast vector. Following the subtraction procedure the subtracted YAC DNA is purified, digested with restriction enzymes and cloned into a plasmid vector (Ostrander et al., 1992). The cloned products of each YAC are screened using a CA(15) oligonucleotide probe (i.e. an oligonucleotide having 15 CA repeats). Each positive clone (i.e. those that contain TG-repeats) is sequenced to identify primers for PCR to genotype the BP-I samples.

An alternative approach, based on using a set of degenerate sequencing primers that anneal directly to the repeat sequence, permitting direct thermal cycle sequencing (Browne & Litt, 1992), can also be used.

Once the candidate region is narrowed to a size of less than about 500 to 1000 Kb, a contiguous array (contig) of clones with smaller inserts than YACs, mainly P1 clones, is developed. P1 clones are phage clones specially designed to accommodate inserts of up to 100 Kb (Shepherd et al., 1994).

DEVELOPMENT OF A PHYSICAL MAP OF THE 5 CM 18PTER REGION

In parallel with the genetic mapping, a physical map of the 5 cM 18pter region is developed. The backbone of this effort is the assembly of contigs of large insert clones. Low resolution contigs for most of the human genome are already available using the YACs developed by CEPH (Cohen et al., 1993). Although these have been individually verified and checked for overlap with other YACs, there is a high rate of chimerism in the YACs and insufficient evidence to definitively confirm the order of the YACs. In addition, because of their large size these YACs are particularly cumbersome to work with. Nevertheless, they provide a useful framework to start constructing high resolution contigs.

Once a candidate region of less than about five cM is delineated, the studies to develop a physical map are commenced. Because of the disadvantages of relying solely on YACs, and because positional cloning is facilitated by the availability of a higher resolution map, contigs are generated using P1 clones once the candidate region is narrowed to less than one Mb, by LD mapping in the expanded population sample using the new markers identified from the YACs.

Once a region of 500-1000 Kb or less is defined, physical mapping and cloning are computed using P1 clones rather than YACs, and P1 contigs over such a region are constructed. The P1s are used to identify additional markers for the further positional cloning steps as well as the screening for rearrangements.

The starting point of contig construction is the microsatellite sequences and non-polymorphic STSs that derive from the few YACs that surround the genetically determined candidate region. These STSs are used to screen the P1 library. The ends of the P1s are cloned using inverse PCR and used to order the P1s relative to each other. Amplification in a new P1 will indicate that it overlaps with the previous one. Fluorescent in situ hybridization (FISH) permits ordering of the majority of the P1s (Pinkel, 1988; Lichter, 1991). The original set of P1s serves as building blocks of the complete contig; each end clone is used to re-screen the library and in this way P1s are added to the map.

From each P1 additional microsatellites are identified as previously described. This allows further reduction of the candidate region. When the region is narrowed to less than one Mb in size, positional cloning efforts are initiated.

FINE MAPPING OF 5CM 18PTER REGION

In order to delineate further regions of BP-I susceptibility within the 5 cM 18pter region, additional unrelated BP-I patients from the CRCV as well as other populations can be diagnosed and genotyped both with the markers described herein as well as additional markers in the 5 cM 18pter region that are known as well those yet to be identified. Additional markers are available from the Cooperative Human Linkage Center (CHLC) public database, from newer Genethon and CHLC maps as they become available (Murray, J.C. et al. (1994) Science 265, 2049-2054, Gyapay, G., et al. (1994) Nature Genet. 7,246-339) and from the public database of the Utah Center for Genome Research (all of which are

incorporated by reference herein). The web addresses for Genethon and CHLC are: Genethon (http://www.genethon.fr/genethon_en.html), CHLC (http://gopher.chlc.org/HomePage.html). These databases are all linked, and one of ordinary skill in the art can readily access the information available from these databases.

The markers shown in **FIG. 6A**, from number 1 to 22 or 23 can be used to genotype the CRCV pedigrees and unrelated BP-I patients described herein as well as other BP-I affected individuals and pedigrees. **See FIG. 6A** (portion of a chromosome 18 map available from the Whitehead Institute, web address: http://133.30.8.1:8080/=@=:www-genome.wi.mit.edu. (incorporated herein by reference)). The fine mapping techniques described herein in conjunction with the teachings regarding the 5 cM 18pter region can be used to narrow the BP-I susceptibility region further.

The following markers (listed in order of occurrence from the telomere towards the centromere) were used to delineate regions of BP-I susceptibility within the 5 cM 18pter region: SAVA5, ca211, ca212, D18S1140, D18S59, ca231, ta201, AT201, ca225, w3442, ca213, ga201, ga203, ca219, D18S1105, ca209, ca202, D18S1146, GATA (referred to in the figures as 166d05) and D18S476. The markers SAVA5, D18S1140, D18S59, ta201, at201, w3442, ga201, ga203, D18S1105, D18S1146, GATA and D18S476 were used in both the haplotype analysis (Figure 8) and the AHR analysis (Figure 11) to delineate the BP-I susceptibility locus to the 500 kb region defined by the markers SAVA5 and ga203 and the 300 kb region defined by D18S1140 and W3422. The other markers were used in both haplotype and the AHR analyses as confirmatory evidence for the localizations. Blood samples from 105 affected individuals were tested for the presence of marker haplotypes and compared to marker haplotypes detected on the non-transmitted chromosome in samples obtained from the parent(s) of the affected individuals when available (71 cases) or to markers detected in samples obtained from a control population of students attending the University of Costa Rica (52 samples). The non-transmitted chromosomes are well matched as controls allowing the affected haplotype of the transmitted chromosome to be more easily discerned than through comparison with data obtained from the general population that may contain individuals who carry the affected haplotype but do not exhibit clinical symptoms of bipolar mood disorder.

Figure 7 provides 18p allele frequencies for disease (aff 105) versus nontransmitted (ntrans) chromosomes and samples from the control population of students (control). The name of each marker used in this study is indicated on the left. The second column of numbers refers to allele length in basepairs. This data provides evidence of over-representation of a particular allele on disease chromosomes.

Figure 8 summarizes the results obtained with affected individuals. The column labelled 18p refers to the patient identifier, and each patient identifier is repeated to indicate results with both copies of chromosome 18. The labels "PANR" and "MANR" refer to the paternal and maternal identifier, respectively, associated with the particular patient, other than 0, 1 and 2, which indicate that parental samples were not available. The allele length (base pairs) is indicated under each marker for a particular patient; the length of the horizontal bar in the figure reflects whether haplotypes are IBD or IBS, with IBD haplotypes with common ancestors having longer bars than randomly inherited IBS haplotypes. To the right of each marker, a "1" indicates that the phase is known, i.e., that it is known whether a particular allele is inherited from the paternal or maternal chromosome, and a "0" indicates that the phase is not known for sure. The determination of phase allows the practitioner to conclude that marker alleles are linked in a haplotype on the same disease causing chromosome.

Figure 9 provides similar data for non-transmitted chromosomes obtained from parental samples. Some individuals exhibited the affected haplotype indicating that the parent was homozygous; however, these regions of identity were typically much shorter than those regions observed in affected individuals, indicating that they were IBS.

Figure 10 similarly provides data for an unscreened population of students from the University of Costa Rica and their parents (52 samples in total). The data demonstrate that there is a lower incidence of the affected haplotype in the general population as compared with Figure 8 and that the affected haplotype is typically shorter similar to the results obtained with non-transmitted chromosomes. However, the results for the general population is less distinctive than that observed for non-transmitted chromosomes in allowing one to map the affected haplotype.

Comparison of the affected haplotype with non-transmitted chromosome markers indicate that the region of maximal sharing between affected individuals occurs between 1140t and w3442 on chromosome 18. This region encompasses approximately 300 kb.

The data was analyzed further using Ancestral Haplotype Reconstruction (AHR), a likelihood method for measuring LD. Data from affected individuals are examined in 2-marker segments. Within each segment, the multinomial likelihood of each of the possible ancestral haplotypes giving rise to the observed sample of disease haplotypes is calculated. This likelihood is calculated assuming some fraction, α, of disease chromosomes are associated with this 2-marker segment, and $(1-\alpha)$ are linked to this segment. These haplotype likelihoods are weighted by the probability of observing that haplotype in the population, and summed to create an overall likelihood for the 2-marker segment. This segment likelihood is compared to the null likelihood, which assumes the disease and markers are unlinked (and therefore $\alpha=0$), and a LOD score is generated. The LOD score is maximized over the parameter α. Details of these calculations are presented in Appendix A. The results of this analysis are shown in Figure 11. The percentages given above the diagonal line demarcated by the filled boxes indicate the percentage of disease chromosomes hypothesized to be true chromosomes from a common founder. For example, 17% of chromosomes obtained from affected individuals have the 18S59 to W3442 region; i.e., as each individual has two chromosome copies, 34% of individuals have this region. The number above each percentage indicates the LOD score. The numbers given below the diagonal line demarcated by the filled boxes indicate the alleles inherited from a common founder, with the number prior to the dash indicating the allele of the marker identified in the column heading and the number following the dash indicating the allele of the marker identified in the row heading. The marker alleles are referred to as follows:

MARKER	#	ALLELE LENGTH
SAVA5	2	229
CA211	3	195
18S1140	2	268
18S59	4	154
18S59	6	158
TA201	2	220
TA201	3	230
CA231	2	186
CA231	4	202
AT201	1	170
AT201	2	178
CA225	1	160
CA225	3	172
W3442	1	10

Blank boxes indicate no positive evidence for linking the indicated region to the affected chromosome.

USE OF P1 CLONES TO IDENTIFY CANDIDATE cDNAs FOR SCREENING FOR MUTATIONS IN THE DNA OF BP-I PATIENTS

The P1 clones described above are used to identify candidate cDNAs. The candidate cDNAs are subsequently screened for mutations in DNA from BP-I patients. From the minimal candidate region defined by genetic mapping experiments a segment is left that is sufficiently large to contain multiple different genes.

IDENTIFICATION OF CODING SEQUENCES

Coding sequences from the surrounding DNA are identified, and these sequences are screened until a probable candidate cDNA is found. Much of the human genome will be sequenced over the next few years, in which case it may become feasible to identify coding sequences through database screening. Candidates may also be identified by scanning

databases consisting of partially sequenced cDNAs (Adams et al., 1991), known as expressed sequence tags, or ESTs. These resources are already largely developed, and include upwards of 100,000 cDNAs, the majority expressed primarily in the brain. It is not yet clear, however, that the complete set of cDNAs will be mapped to specific chromosomal locations in the near future, and that their data will soon be made publicly available. The database can be used to identify all cDNAs that map to the minimal candidate region for BP-I. These cDNAs are then used as probes to hybridize to the P1 contig, and new microsatellites are isolated, which are used to genotype the "LD" sample. Maximal linkage disequilibrium in the vicinity of one or two cDNAs is identified. These cDNAs are the first ones used to screen patient DNA for mutations. Database screening has already been used to identify a gene responsible for familial colon cancer (Papadopolous et al., 1993).

Coding sequences are also identified by exon amplification (Duyk et al., 1990; Buckler et al., 1991). Exon amplification targets exons in genomic DNA by identifying the consensus splice sequences that flank exon-intron boundaries. Briefly, exons are trapped in the process of cloning genomic DNA (e.g. from P1s) into an expression vector (Zhang et al., 1994). These clones are transfected into COS cells, RT-PCR is performed on total or cytoplasmic RNA isolated from the COS cells using primers that are complementary to the splicing vector. Exon amplification is tedious but routine; for example, the system developed by Buckler et al. (1991). This method is probably preferable to another widely used approach, direct selection, which involves screening cDNAs using large insert clone contigs, with several steps to maximize the efficiency of hybridization and recovery of the appropriate hybrid (Lovett et al., 1991). Although direct selection is more efficient than exon amplification (Del Mastro et al., 1994), it may not be practical as it depends on the candidate cDNA being expressed in the tissue from which the cDNA library was made; there is no prior information to indicate the tissue or developmental stage in which BP-I genes would be expressed.

Once cDNAs are identified the most plausible candidates are screened by direct sequencing, SSCP or using chemical cleavage assays (Cotton et al. 1988).

The data are also evaluated for clues to the possible identity or mode of action of BP-I mutations. For example, it is known that trinucleotide repeat expansion is associated with

the phenomenon of anticipation, or the tendency for a phenotype to become more severe and display an earlier age of onset in the lower generations of a pedigree (Ballabio, 1993). Several investigators have suggested that segregation patterns of BP-I are consistent with anticipation (McInnis et al., 1993; Nylander et al., 1994). The apparent transmission of BP-I, in association with the conserved 18q23 haplotype is constant with anticipation. Therefore, once the candidate region is narrowed to its minimal extent, the P1 clones are screened using trinucleotide repeat oligonucleotides (Hummerich et al., 1994). A PCR assay is developed and patient DNAs are screened for expanded alleles.

Genetic and physical data help to map the bipolar mood disorder gene to the 5 cM 18pter region of chromosome 18. New markers from this region are tested in order to locate the bipolar mood disorder gene in a region small enough to provide higher quality genetic tests for bipolar mood disorder, and to specifically find the mutated gene. Narrowing down the region in which the gene is located will lead to sequencing of the bipolar mood disorder gene as well as cloning thereof. Further genetic analysis employing, for example, new polymorphisms flanking D18S59 and D18S476 as well as the use of cosmids, yeast artificial chromosome (YAC) clones, or mixtures thereof, are employed in the narrowing down process. The next step in narrowing down the candidate region includes cloning of the chromosomal region 18pter including proximal and distal markers in a contig formed by overlapping cosmids and YACS. Subsequent subcloning in cosmids, plasmids or phages will generate additional probes for more detailed mapping.

The next step of cloning the gene involves exon trapping, screening of cDNA libraries, Northern blots or rt PCR (reverse transcriptase PCR) of samples from affected and unaffected individuals, direct sequencing of exons or testing exons by SSCP (single strand conformation polymorphism), RNase protection or chemical cleavage.

Flanking markers on both sides of the bipolar mood disorder gene combined with D18S59 and D18S476 or a number of well-positioned markers that cover the chromosomal region (5 cM 18pter) carrying the disease gene, can give a high probability of affected or non-affected chromosomes in the range of 80-90% accuracy, depending on the informativeness of the markers used and their distance from the disease gene. Using current markers linked to bipolar mood disorder, and assuming closer flanking markers will be

identified, a genetic test for families with bipolar mood disorder will be for diagnosis in conjunction with clinical evaluation, screening of risk and carrier testing in healthy siblings. In the future, subsequent delineation of closely linked markers which may show strong disequilibrium with the disorder, or identification of the defective gene, could allow screening of the entire at-risk population to identify carriers, and provide improved treatments.

TREATMENT OF BP-I PATIENTS USING GENOTYPE DATA

Using the fine mapping techniques described herein, BP-I susceptibility loci or genes in the 5 cM 18pter region in particular in the region #1 between SAVA5 and ga203, are identified and used to genotype patients diagnosed phenotypically with BP-I. Preferably, markers in the roughly 500 kb region defined by SAVA5 and ga203, inclusive, are used. More preferably, markers in either the region defined by D18S59 and w3422, inclusive, are used.

Genotyping with the markers described herein as well as additional markers permits confirmation of phenotypic BP-I diagnoses or assist with ambiguous clinical phenotypes which make it difficult to distinguish between BP-I and other possible psychiatric illnesses. A patient's genotype in the 5 cM 18pter region is determined and compared with previously determined genotypes of other individuals previously diagnosed with BP-I. Once an individual is genotyped as having a BP-I susceptibility locus in the 5 cM 18pter region, the individual is treated with any of the known methods effective in treating at least certain individuals affected with BP-I, such as the administration of lithium salts, carbamazepine or valproic acid.

Studies are conducted correlating effective treatments with BP-I genotypes in the 5 cM 18pter region to determine the most effective treatments for particular genotypes. BP-I patients can then be genotyped in the 5 cM 18pter region and the statistically most effective treatment can be determined as a first course of therapy.

All publications and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference.

The invention now being fully described, it will be apparent to one of ordinary skill in the art that many changes and modifications can be made thereto without departing from the spirit or scope of the appended claims.

Appendix A

Consider the original mutation to have occurred on a chromosomal segment between two markers A and B. At the time the mutation was introduced, there were n_a alleles at marker A and n_b alleles at marker B. On the chromosome containing the disease mutation both marker A and marker B carried allele X. The probability that after g generations an affected individual carrying the original disease mutation would still have allele X at markers A and B is:

$$(1-\theta_1)^g(1-\theta_2)^g + (1-\theta_1)^g(1-(1-\theta_2)^g)f(X_B) + (1-(1-\theta_1)^g)(1-\theta_2)^g f(X_A) +$$

eq (1)
$$(1-(1-\theta_1)^g)(1-(1-\theta_2)^g)f(X_A)f(X_B)$$

where θ_1 is the recombination fraction between disease and marker A, θ_2 is the recombination fraction between disease and marker B, g is the number of generations since founding (i.e. since the mutation was introduced into the population), $f(X_A)$ is the population frequency of the X-allele at marker A and $f(X_B)$ is the population frequency of the X-allele at marker B. This equation includes terms for the possibility of recombination between the markers and the disease locus, with the X-allele at the markers then being identical by state (IBS) rather than IBD. The probabilities of an affected individual with the original mutation having other haplotypes can be formulated similarly. The probability of having allele Z at marker B (where Z is any allele at marker B besides X) would be:

$$(1-\theta_1)^g(1-(1-\theta_2)^g)f(Z_B) + (1-(1-\theta_1)^g)(1-(1-\theta_2)^g)f(X_A)f(Z_B)$$

eq (2)

where $f(Z_B)$ is the frequency of allele Z at marker B in the population. The probability of having allele Z at marker A (where Z is any allele at marker B besides X) would be :

$$(1-\theta_2)^g(1-(1-\theta_1)^g)f(Z_A) + (1-(1-\theta_1)^g)(1-(1-\theta_2)^g)f(X_B)f(Z_A)$$

eq (3)

where $f(Z_A)$ is the frequency of allele Z at marker A in the population. Finally, the probability of having allele Z at both markers A and B would be:

$$(1-(1-\theta_1)^g)(1-(1-\theta_2)^g)f(Z_A)f(Z_B)$$

eq (4)

These probabilities assume (1) no interference in recombination and (2) the same marker alleles are present now as were present *g* generations ago, in similar frequencies. If, for example, marker A has n_a alleles and marker B has n_b alleles, then these probabilities form a (n_a)-(n_b) by (n_a)-(n_b) transition matrix, with row *i* containing the probabilities that founder haplotype *i* gave rise to each of the (n_a)-(n_b) different haplotypes in *g* generations. The rows of this transition matrix sum to 1.

In simulations, the haplotype frequencies in the disease population were formulated using these transition probabilities, assuming the disease arose on a haplotype with the "1" allele at each of the two markers.

Once these transition probabilities are estimated, the likelihood of a particular founder chromosome giving rise to the observed sample of disease haplotypes in g generations is easily estimated. For example, if one assumed that the disease mutation arose on a chromosome with the X-allele at both markers, the likelihood ($L_{X\text{-}X}$) that this chromosome was the founder of the present-day sampled disease chromosomes is given by the multinomial:

$$L_{X-X} = \prod_{i=1}^{K} (p_{X-X,i})^{Y_i}$$

eq (5)

where i indexes the K potential haplotypes for the two markers ($K=(n_a)(n_b)$), $p_{X\text{-}X,i}$ is the probability that the ancestral disease chromosome with the X-allele at both markers gave rise to a haplotype of type i in g generations (taken from the transition matrix), and Y_i is the observed number of haplotype i in the sample ($\Sigma_i(Y_i)$=the number of chromosomes in the sample to be analyzed). The likelihood in eq (5) assumes that all affected individuals are independent. While, after many generations of separation from a common ancestor one might consider these

individuals to be independent, they are in fact related through a complex and unknown pedigree. The simplification of considering individuals as independent makes the likelihood much more tractable to compute.

The K likelihoods are then summed, and weighted by the probability of observing that particular haplotype in the population to produce an overall likelihood for the 2-marker segment:

eq (6)
$$L = \sum_{i=1}^{K} f_i L_i$$

where f_i is the frequency of haplotype i in the population. This overall likelihood calculation parallels the approach taken by Terwilliger (1995, eq (2)). The haplotype frequencies are estimated from the sample of normal chromosomes. In the event that the haplotype resulting in the largest contribution to the overall likelihood in eq (6) is not observed in the normal sample, the upper 95% confidence interval for this frequency is used, and the remaining haplotype frequencies rescaled accordingly.

This overall likelihood is compared to the null likelihood, which is generated in exactly the same manner, except that it is assumed the markers were unlinked to

the disease locus ($\theta_1=\theta_2=0.5$ in, for example, eqs (1-4)). The $\log_{10}$ of this likelihood ratio is a LOD score. One might consider to use in the null likelihood transition probabilities calculated under the assumption of linkage equilibrium. Under this null the cells of the transition matrix are computed by multiplication of allele frequencies, assuming independence of marker loci. These two forms of the null likelihood are equivalent in value for *g* of approximately 20 or greater, and for $g<20$ the values are nearly equivalent.

Because θ_1 and θ_2 are obviously unknown, the putative disease locus is set to be in the middle of the segment and therefore θ_1 and θ_2 are one-half the genetic distance (converted to recombination fraction by the Haldane mapping function, (Ott 1991)) between the two marker loci forming the segment. In fact, one could estimate θ_1 and θ_2, or their ratio, and the method could easily be modified to do so, however for our purposes finding a linked segment is suitable.

This basic procedure has been modified to deal with heterogeneity in the sample of disease chromosomes. Not all chromosomes in the disease sample may be true disease chromosomes from a common founder. Individuals heterozygous for the disease mutation will add one chromosome to the disease sample that will not be a true disease chromosome. Additionally, affected individuals not linked to the

particular chromosomal segment being analyzed (either because they are phenocopies or because of locus heterogeneity) will contribute two chromosomes to the affected sample that do not harbor this disease locus. When the null hypothesis of no linkage is not true, some fraction, α, of the chromosomes in the disease sample will associated with this chromosomal segment, and $(1-\alpha)$ will not be associated. We decided to examine α in steps of 0.1, from 1.0 to 0.0, and for each step in α produce a new transition matrix under the alternative hypothesis and calculate a LOD score. If we call the transition matrix calculated under the alternative hypothesis (where the disease locus is hypothesized to be in the middle of the 2-marker segment) T_a and call the transition matrix calculated under the null hypothesis (where the disease locus is unlinked to the 2-marker segment) T_n, then a new transition matrix for the alternative hypothesis is calculated as:

$$T^*_a = \alpha T_a + (1-\alpha) T_n$$

eq (7)

The transition matrix under the null uses $\alpha=0$. The LOD score is then maximized over the one parameter α.